Implementation of Automated Industry-Specific Identification of a Grant Application Based on Tokenization with a Predetermined Weight

Andrey V. SIROTKIN

Ph.D. (Engineering), Associate Professor
Associate Professor of the Department of Mathematics and Computer Science
Institute of Digital Technologies and Economics
North-Eastern State University
4, Commune Str., Magadan, 685000, Russia
E-mail: avsl100@yandex.ru

Vadim K. KOPCHENKO

4th year student
Institute of Digital Technologies and Economics
North-Eastern State University
2, Polar str., Magadan, 685000, Russia
E-mail: kopchenko49@yahoo.com

Abstract

The article discusses the process automation of industry-specific identification of an application for project financing in an automated grant distribution system (ASGD) using tokenization of the application text and weighting of many predefined terms. The study aim is to develop a module for analyzing the industry sector of an application. The research methodology is composed of systems analysis and mathematical modeling methods. Existing developments are analyzed. Functional models for canonicalizing the application text have been developed, allowing the use of terminological analysis methods to study the document. A database model has been developed for organizing the storage of information about the application, key terms and their weighting factors. An example of the operation of this segment of the system. The module is implemented as a web application in PHP using the Apache HTTP-server, MySQL database management system and RedBeanPHP object-relational mapping library.

Keywords: grant; automation; system; distribution; programming; token; lemma; normalization; documentary frequency.

Introduction

Modern automated data processing systems solve a wide range of tasks, which have recently actively covered decision-making processes. An example of such a task can be a grant when there is an application for project financing, and depending on the industry, novelty and a number of additional criteria, a decision is made whether to satisfy this application or not. In such activities, the practice of attracting experts (Melnik, 2014, 2017, 2018), despite the fact that for this class of tasks there is no widespread coverage in scientific journalism of information on the development and use of automated systems. Attempts were made to conceptual design an ASGD (Kopchenko, 2019; Sirotkin, 2019). ASGD was considered from the queuing system point of view, where the decision-making task will be to find such a solution q_j from the set of alternative options for which the value of the generalized indicator (1) is the most extreme (2) (Kopchenko, 2019):

$$\varphi = F(N) \rightarrow \text{ext}, \quad (1)$$

N – a set of indicators, F – convolution function of particular criteria, the appearance of which is not defined due to further work of the authors on increasing the universality of the solution.

$$q_j = Arg\ ext(\varphi), \quad (2)$$

$q_j \in Q$, Q – a set of alternative solutions, $j \in \{1, \overline{Q}\}$.

The ASGD is considered based on industry patterns and knowledge bases (Sirotkin, 2019). The organizational-process model of such a system with a TO-BE timestamp can be represented as follows:

$$\theta_{TO-BE} = \langle \Psi, P_p, M, Q_p, Q_M, O_M \rangle, \quad (3)$$

Ψ - a set of applications, P_p - a lot of preliminary assessment procedures (for compliance with the requirements), M - a set of expert assessment procedures, Q_p - a set of criteria for a preliminary assessment (for compliance), Q_M - a set of criteria for an expert assessment, O_M - a set of numerical evaluations. Figure 1 shows the context diagram of the operation of ASGD. The purpose of the application review process is to directly finance the project.

Figure 1. ASGD context diagram

Despite the scope of the studies examined, the question of industry-specific identification of the application remains open with the aim of determining the criteria for automated assessment and choosing the appropriate knowledge base. One of the best-known text rubricating methods is to weight the terms of a document. A vivid example is the calculation of the TF-IDF measure, according to which the weight of a word is proportional to the frequency of use of this word in the document and inversely proportional to the frequency of use of the word in all documents of the collection (Salton, McGill, 1983). The formula for calculating the measure is the product of the word frequency by the inverse frequency of the document. The word frequency is calculated using the following formula:

$$tf(t, d) = \frac{n_t}{\sum_k n_k}, \quad (4)$$

t is a term; d is the document; n_i is the number of occurrences of the term in the document; $\sum_k n_k$ – is the total number of terms in the document. The reverse frequency of the document is calculated by the formula:

$$idf(t, D) = \log \frac{|D|}{|d_i \in D \mid t \in d_i|}, \quad (5)$$

$|D|$ is the number of documents in the collection, $|d_i \in D \mid t \in d_i|$ – is the number of documents from the D collection in which t occurs (if $n_t \neq 0$). Accounting for IDF reduces the weight of commonly used words. The calculation formula for the TF-IDF measure is as follows:

$$tf - idf(t, d, D) = tf(t, d) \times idf(t, D). \quad (6)$$

In this paper, we consider a special case of the TF-IDF measure, when the composition of key terms and their corresponding weights is predetermined for each industry in advance, that is:

$$\forall o_i \in I \exists w_i = const, \quad (7)$$

o is the term of the industry, I is the industry, w is the weight of the term, $i = |I|$. The following model can express weights:

$$w_i \in W, \quad (8)$$

W is a set of weights, $i \in [1, |W|]$. Each weight is set in accordance with an expert assessment of the importance of the term in industry identification. The assessment model can be represented as follows:

$$w = F(T, O, P_o, C), \quad (9)$$

T - term, O - sector, P_o - estimation, C - expert assessment of importance.

The aim of the study is to develop a module for industry-specific application identification in an automated grant distribution system based on tokenization with a predetermined weight.

The study objectives are:

1. Publications analysis on the research topic.
2. Mathematical model description of industry identification.
3. Database development a for an automated grant distribution system, considering the selected application identification method.
4. Development of industry identification modules for the system.

Methods

The study is based on the use of system analysis methods and mathematical modeling. The practical part of the study is based on the theory of databases and programming languages.

Results

As a sample application for a grant, a form consisting of 10 fields is used. The project name, thematic area, project implementation goals, project objectives, innovative aspects, justification of the need for Research and Development, technical parameters, design requirements, scope and description. Each field contains a text description, that is, the application can be represented as a set of words whose power is equal to their number (Sendhilkumar, Nachiyar, Mahalakshmi, 2013):

$$Z = \{z_1, z_2, z_3, \dots, z_n\}, |Z| = n, \quad (8)$$

Z – a set of text words, z_i i -th word of the text, n – the number of words in the text. In turn, the entire set of text words can be represented as a union of k subsets of different parts of speech, while each word of the text can be assigned to one of these subsets:

$$Z = \bigcup_{j=1}^k C_j; z_i \in C_j; i = \overline{1, n}; j = \overline{1, k}, \quad (9)$$

C_j – a subset of the words of the j -th part of speech, k – the number of parts of speech in the language. The initial task is to extract words from the application text (tokenization) and bring each token to the normal form of the lemma. Lemmatization is required to simplify the weighing of

tokens during industry identification, since in the terminology of computer linguistics each word form of the same word is a new token. Each part of speech has its own order of reduction to the main form. For example, for nouns it is the nominative case and being singular, for adjectives it is the nominative case, being singular and masculine gender. The application identification process is presented on the *IDEF0* chart (Figure 1). According to GOST R 50.1.028-2001 (Gosstandart of Russia, 2003), the following glossary applies to the diagrams of this notation:

1. Function - an activity, process or transformation (modeled by the *IDEF0* block), identified by a verb or verb form that describes what needs to be done. In our case, these are activities such as tokenization, lemmatization and weighting, which implement the main identification process
2. Arrow - A directional line consisting of one or more segments that simulates an open channel or channel transmitting data or material objects from a source (i.e. the starting point of the arrow) to the consumer (end point with a 'tip'). In our case, the most important arrows are the flows of tokens and lemmas used in the identification process.
3. Control arrow - A class of arrows that indicate controls in *IDEF0*, that is, the conditions under which the output of the block will be correct. Data or objects modeled as controls can be transformed by a function that creates the corresponding output. The control arrows are associated with the top side of the *IDEF0* block. In our case, management is carried out using dictionaries.
4. Arrow mechanism - A class of arrows that display the mechanisms *IDEF0*, that is, the means used to perform the function; includes a special case of a call arrow. The arrows of the mechanisms are associated with the lower side of the *IDEF0* block. In our case, the mechanism are computer technologies and identification procedures.
5. Input arrow - A class of arrows representing the input of an *IDEF0* block, that is, data or material objects that are converted by the function into output. Input arrows are associated with the left side of the *IDEF0* block. In our case, an application for financial and credit support is submitted to the system entrance.
6. Output arrow - A class of arrows representing the output of an *IDEF0* block, that is, data or material objects produced by a function. The output arrows are associated with the right side of the *IDEF0* block. The result of the system operation can be considered a decision to satisfy the application or to refuse it.

Using auxiliary procedures, the system conducts industry identification using data from the dictionary. The input to the identification process is the full text of the application. Further, using transformations using dictionaries, the industry affiliation of the project is determined.

Figure 2. Context Chart ‘Conduct Industry Identification’

The decomposition of the context diagram is presented in figure 3. Based on this model of business processes, the following reasoning can be presented. The input of the 'Tokenization' job is submitted with the full text of the application, using the procedures defined during the development of the automated system and the data of the stop-word dictionary, the text is split applications for individual terms, each term is reduced to lower case. Further, the set of tokens obtained from the application is processed by the work ‘Lemmatization’, where, using a dictionary from the free office suite OpenOffice.org and the phpMorpho library, each term undergoes normalization, that is, reduction to the main form of the word. Upon completion of the lemmatization process, a set of lemmas is served as an input to the work ‘Weighing’, where for each branch of knowledge a set of keywords and their corresponding weights are loaded, after which nested loops search for tags across a set of lemmas.

Figure 3. Decomposition of the context diagram ‘Conduct industry identification’

Each match increases the membership index of the application to a particular branch of knowledge by the value of tag weight to which a match was found. The industry with the highest indicator value is the most suitable. The result of the work is an application with recognized industry affiliation.

To store the industry vocabulary and the application itself, the ASGD database is developed, the logical model of which is presented in Figure 4.

Figure 4. Logical model of ASGD database

The ASGD database consists of four entities, i.e. ‘Client’, ‘Requirement’, ‘Industry’ and ‘Industry terms’. The entity ‘Client’ contains information about the registered ASGD user, is related ‘one to

many’ with the entity ‘Requirement’, which contains information about the object of co-financing. The entity ‘Industry’ contains a list of industries for categorizing the application, linked to the entity ‘Industry terms, which contains information on terms and weights belonging to a particular industry.

The mathematical formulation of the tokenization problem can be defined as follows: each element z_i of the set of terms of the application Z consists of the letter set of the Russian alphabet B and the set of s characters. Punctuation marks, numbers and diacritical characters may be elements of this set:

$$z_i = \{B_1, s_1, \dots, B_m, s_m\}, \quad (10)$$

m – the number of all characters in one term application. Then each t_i token of the application will take the following form:

$$t_i = \{b_{1a}, b_{2a}, \dots, b_{ja}\}, \quad (11)$$

j – the number of letters in the term z_i , a – number of characters in the Russian alphabet (Petrov, 2017).

According to studies, words of the Russian language are divided into many stop words (12), which includes prepositions, conjunctions, particles, interjections and pronouns, and many word forms (13), that is, the application term cannot belong to both sets at once (14) (Borodin, 2008).

$$ZS = \{zs_1, zs_2, \dots, zs_x\}, x = |ZS|, \quad (12)$$

$$ZF = \{zf_1, zf_2, \dots, zf_y\}, y = |ZF|, \quad (13)$$

$$(t_i \in ZS) \wedge (t_i \in ZF). \quad (14)$$

Under the condition that $t_i \in ZF$, the lemmatization problem can be represented as a functional correspondence (15), the result of which is a set of normalized words.

$$F = \langle ZF, ZN \rangle, \quad (15)$$

ZN – a set of lemmatized tokens, $ZN = \{zn_1, zn_2, \dots, zn_{|ZN|}\}$, zn_i – lemma.

The final step in the process of industry identification is the search for each industry knowledge of the terms corresponding to it in the normalized application text. Let G be the set of branches of knowledge, $G = \{g_1, g_2, \dots, g_c\}$, $c = |G|$, then:

$$\forall g \exists w \neq 0, \quad (16)$$

w – term weight. Then the value r of the correspondence of the g -th branch of knowledge is determined as follows:

$$r = \max \sum_{i=1}^{|G|} w_{ij}, \quad (17)$$

w_{ij} – weight of the j -th term in the i -th industry, $j = \overline{1, |g|}$.

This model was implemented in PHP in the industry identification module of the automated grant distribution system using the RedBeanPHP object-relational mapping library and phpMorpho morphological analysis library (Prettyman, 2016; Mitchell, 2016). The phpMorpho dictionary contains the basics of words, the rules for changing them and the grammatical metadata necessary for correct lemmatization. The dictionary is binary compatible between different platforms, which simplifies system migration.

Discussion

The results obtained allow us to conclude that the grant distribution system is ready for further modernization and implementation of new functional modules, and emphasize the need for a mathematical definition of the competitive procedure (Sirotkin, 2019). Active development includes a module for determining the scientific novelty of an application based on an analysis of the relevance of search results for the purpose of the project, as well as a module for verifying the application text for incorrect borrowing, otherwise called plagiarism, which will be implemented using the shingles algorithm, which involves 4 stages (figure 4) (Borodin, 2008). At the first stage of work, the text is canonized, that is, the reference text and the text being checked are cleared of stop words, numbers, punctuation characters, all words are converted to a single register (these functions are already implemented in the automated grant distribution system in the industry identification module). Splitting texts into shingles (stage 2) is a division of the source text into ordered sequences of words of arbitrary size. At the third stage, for each shingle, its hash sum is found. The final step is to compare the hash sums. It is possible to compare each pair of these values, however, to improve performance, samples of the values of the resulting keys are usually compared (for example, only those that are divided by 25) (Broder, 2000). In general, the shingles algorithm is considered not the most reliable way to check the text for borrowing, as changing just one letter generates a completely new hash. Nevertheless, high-quality canonicalization of the text by the tools discussed in this article will help minimize the risk of missing a duplicate. The operation of the shingles algorithm is illustrated in Figure 5.

Figure 5. Shingles algorithm operation

Conclusion

A module for the industrial identification of a grant application has been developed and tested. The module is based on storing in a database for each industry a set of weighted key terms, which are searched in the canonized application text. A model of the shingle algorithm was also presented, the implementation of which was postponed for further research. During the development, text

normalization methods, such as tokenization and lemmatization, were studied; these procedures were also introduced into the automated system. Further research will be aimed at a relevant analysis of the novelty of the application and the type of draw function, which is currently not defined because all possible sets of indicators for all cases of decision-making are not fully formulated.

References

- Borodin, O.N. (2008). Extracting information from full-text data sources. *Bulletin of the Moscow State Agroengineering University. Vestnik FGOU VO MGA*, 1(20). Retrieved from: <https://cyberleninka.ru/article/n/izvlechenie-informatsii-iz-polnotekstovyyh-istochnikov-dannyh> (Accessed on November 4, 2019).
- Broder, A. (2000). Identifying and Filtering Near-Duplicate Documents. *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, 1-10. 10.1007/3-540-45123-4_1.
- Gosstandart of Russia. (2003). P 50.1.028-2001. Information technology support product life cycle. Methodology of functional modeling. Gosstandart of Russia. Moscow: Publishing house of standards, 4-6.
- Kopchenko, V.K. (2019). The use of artificial intelligence systems for the distribution of grants to entrepreneurs of the Magadan region. *Student electronic scientific journal*, 11(55), 20–24.
- Melnik, P.B. (2014). The mathematical formulation of the task of forming the roster of experts. *Innovation and Expertise*, 2(13), 69–81.
- Melnik, P.B. (2017). Methodology of forming expert pools and groups for conducting expert analytical studies. *Innovation and Expertise*, 1(19), 39–54.
- Melnik, P.B. (2018). The roster of experts as a queuing system: model and parameters of the incoming flow of applications. *Innovation and Expertise*, 1(22), 67–78.
- Mitchell, L.J. (2016). PHP Web Services: APIs for the Modern Web. Sebastopol: O'Reilly.
- Petrov, A.S. (2017). A mathematical model of a Russian-language text document for solving the problem of automatically extracting terms from a text. *Bulletin of Voronezh State University*, 3, 195–203.
- Prettyman, S. (2016). Learn PHP 7: Object Oriented Modular Programming using HTML5, CSS3, JavaScript, XML, JSON, and MySQL. New York City: Apress.
- Salton, G., MacGill, M. J. (1983). Introduction to modern information retrieval. N. Y. McGraw-Hill.
- Sendhilkumar, S., Nachiyar, N., Mahalakshmi, G.S. (2013). Novelty Detection via Topic Modeling in Research Articles. *Computer Science & Information Technology*, 3, 401-410. DOI: 10.5121/csit.2013.3542.
- Sirotkin, A.V. (2019). Sector-specific identification of applications in an automated expert grant distribution system. *Modern high technology*, 7, 99–103.
- Sirotkin, A.V. (2019). The concept of developing an automated grant distribution system. *XXI International Scientific and Practical Conference 'EurasiaScience'*, 107–109.